Categorical Variables

Visualizing and Describing

Grinnell College

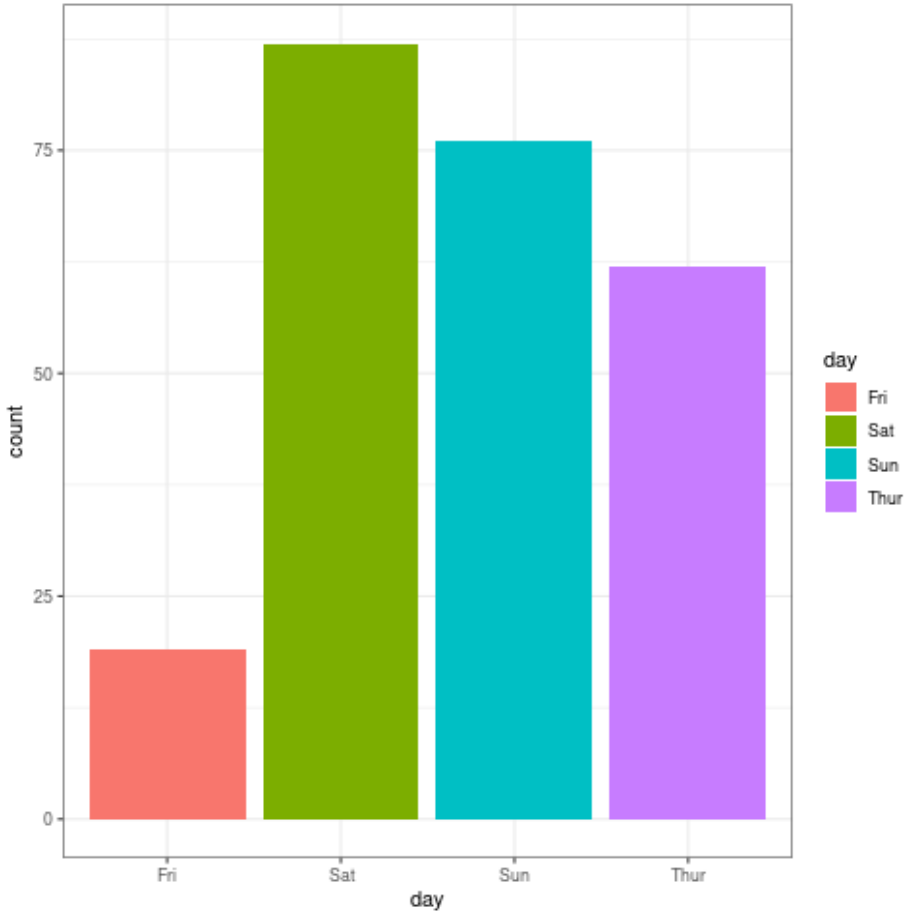
Review – Distribution

In order to better understand patterns in our data, we will often combine graphics with short descriptions of what we see. A term we will use often:

The **distribution** of a variable refers to how frequently certain values of that variable show up in our data

One Categorical Variable → Bar Chart

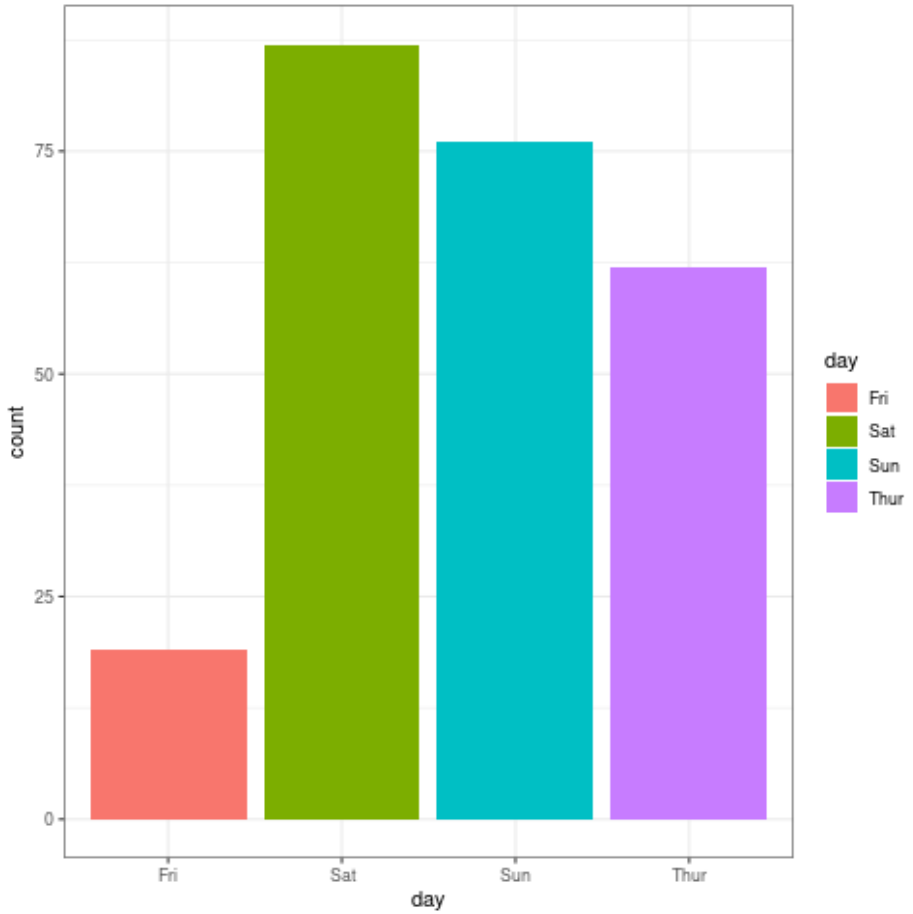
When we have one categorical variable, a *barchart* is often used to tally the frequencies (counts) of that categorical variable



- Notice anything about order?

One Categorical Variable → Bar Chart

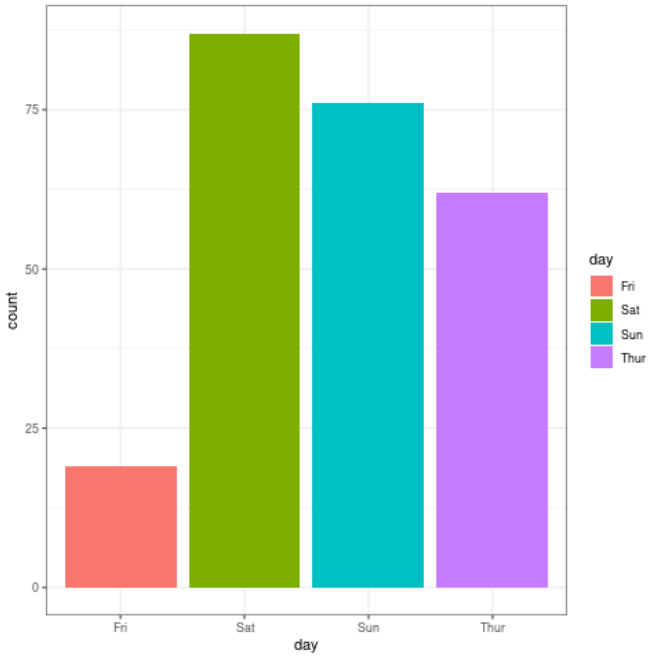
To describe the distribution of a categorical variable, we need to talk about how likely each category is, and mention the most and least likely categories (helpful to include supporting values)



Distribution of customers?

Percentages

A percentage is a ratio of something out of 100. This is used to give us an idea of how often something comes up.



Number of Customers

Friday: 19

Saturday: 87

Sunday: 76

Thursday: 62

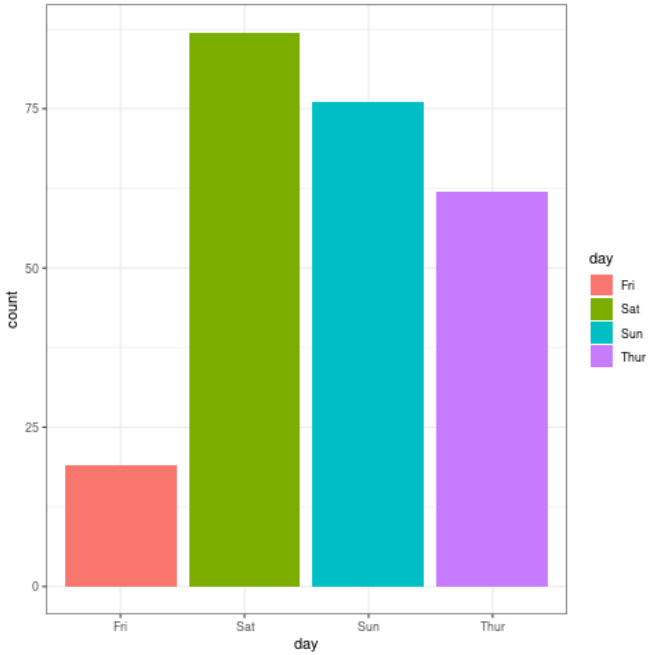
Total: 244

Percentage of the waiter's customers that show up on Friday

- $\frac{19}{244} = .078 = 7.8\%$

Percentages

We can work backwards with percentages to find out how many customers came in on Fridays too.



Number of Customers

Friday: 19

Saturday: 87

Sunday: 76

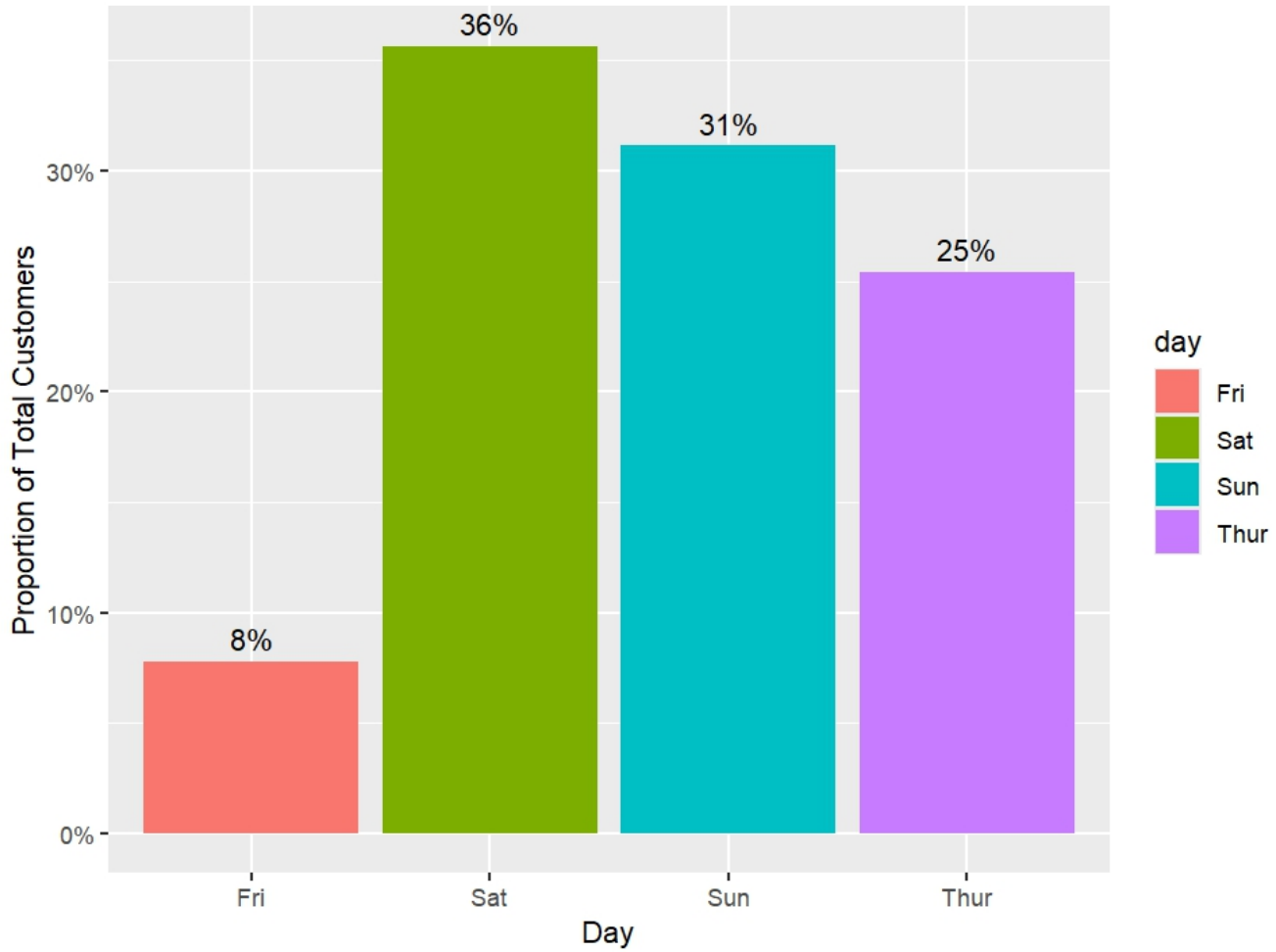
Thursday: 62

Total: 244

Percentage of Friday customers $\times$ 'total customers' = Friday customers

- $7.8\% \times 244 = .078 \times 244 = 19$ (after rounding)

Bar Chart (Relative Frequency)



We could display proportions or percentages directly instead of counts.

- displaying count/prop/percent above bar helps readability
- even so, I may sometimes leave them off to practice estimating by eye

Review – Association

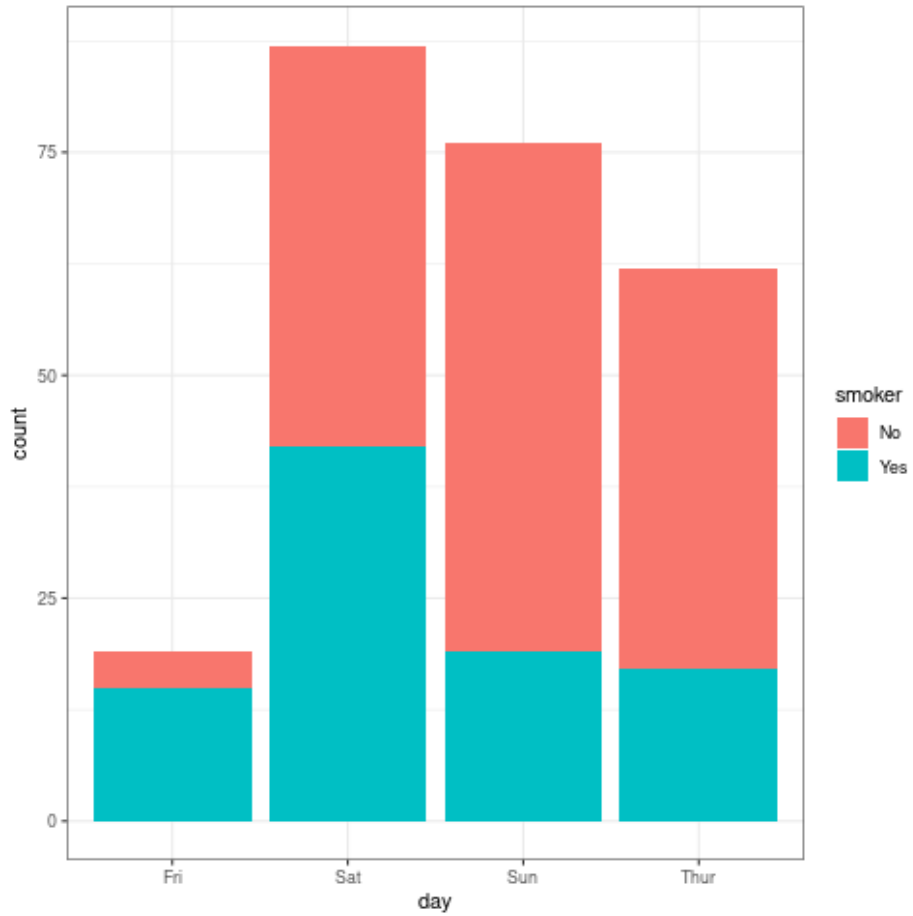
It is very common for us to try to find a relationship between two (or more) variables

- When there seems to be some connection between two variables (knowing about one variable tells us about the other), we say they are **associated**.
- If there does not seem to be a relationship between the variables, we say they are **independent**.

NOTE: this does not always mean that one variable is causing a change in the other

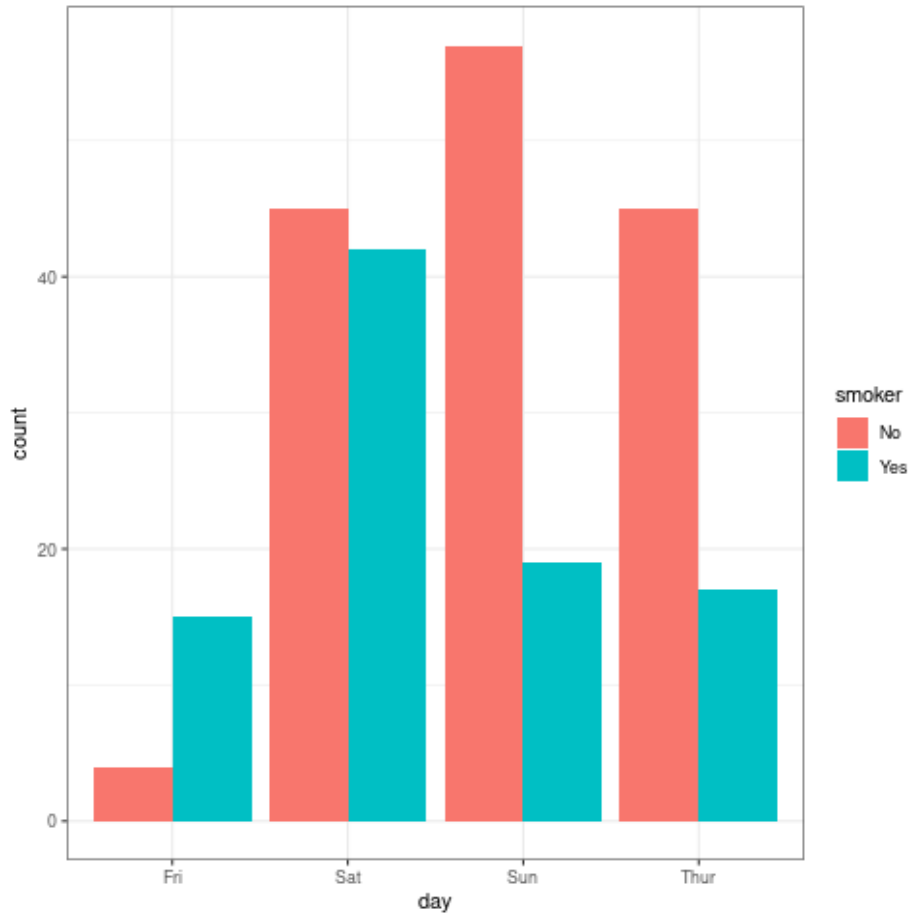
Categorical + Categorical → Stacked Bar

The first type of bivariate bar chart is known as a **stacked bar chart**, which allows us to break down one variable in terms of another. Here, we consider if any smokers were included in the party



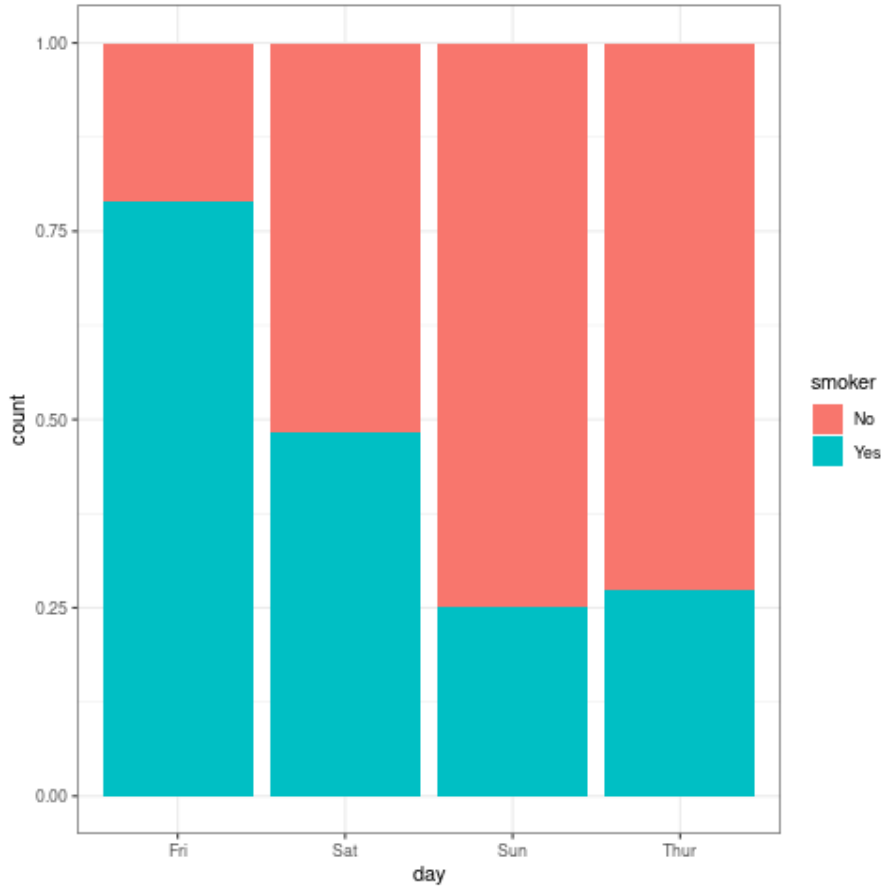
Categorical + Categorical → Dodge Bar

The second type of bivariate bar chart is known as a **dodged bar chart**, which presents both variables alongside one another. This makes comparing within groups much simpler



Categorical + Categorical → Filled Bar

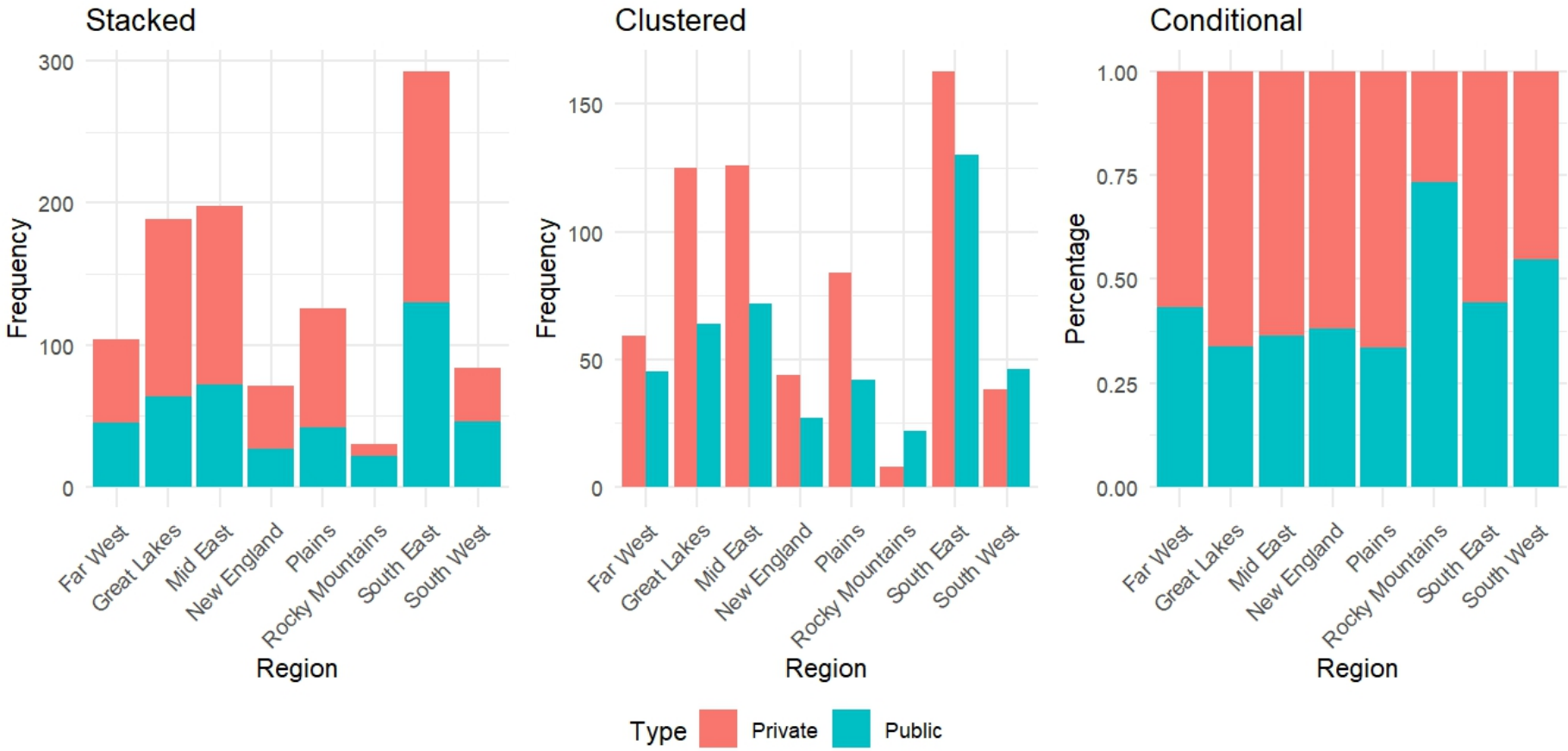
The last type of bivariate bar chart is known as a **filled bar chart**, offering proportions. Although we lose absolute counts, we can now see relative frequencies within each group



- 'GGplot2' defaults to "count" on the y-axis: not good!

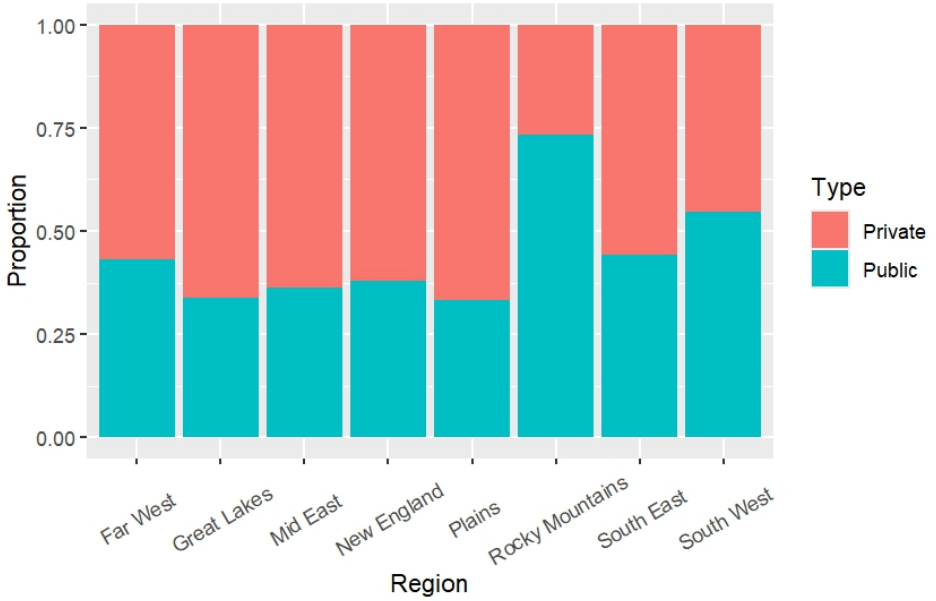
Bivariate Bar Charts

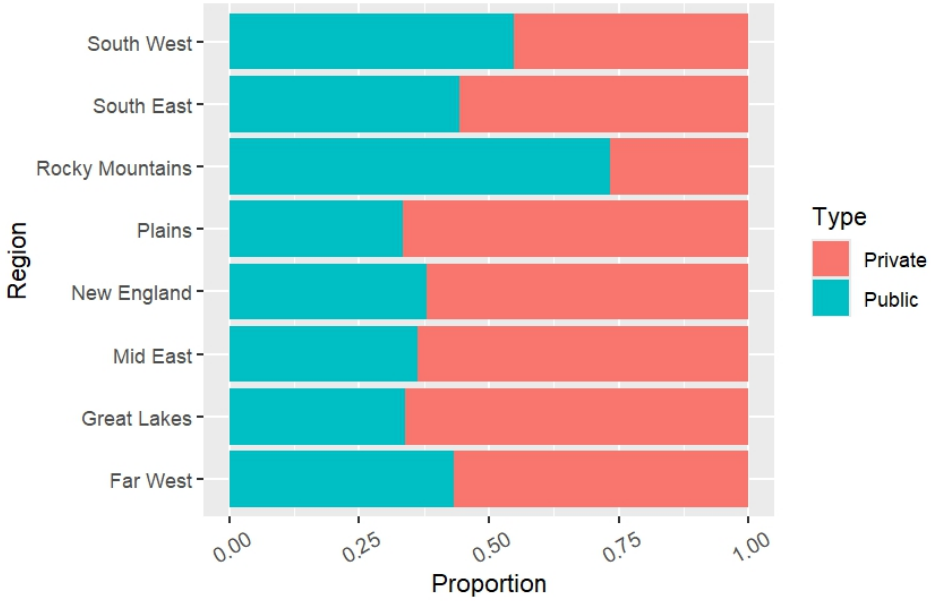
Back to the college data. Are the variables "Region" and "Type" associated? Which bar chart is most helpful?



Order of Axes

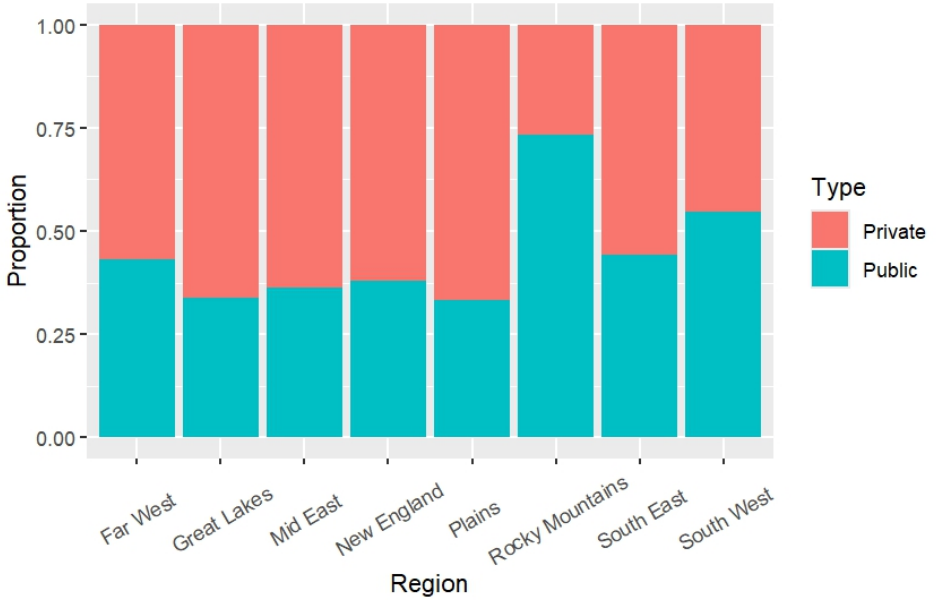
You can change around the x- and y- axes when making graphs. Some people read one or the other more easily. You may have a personal preference for aesthetics and can choose which to use later on.

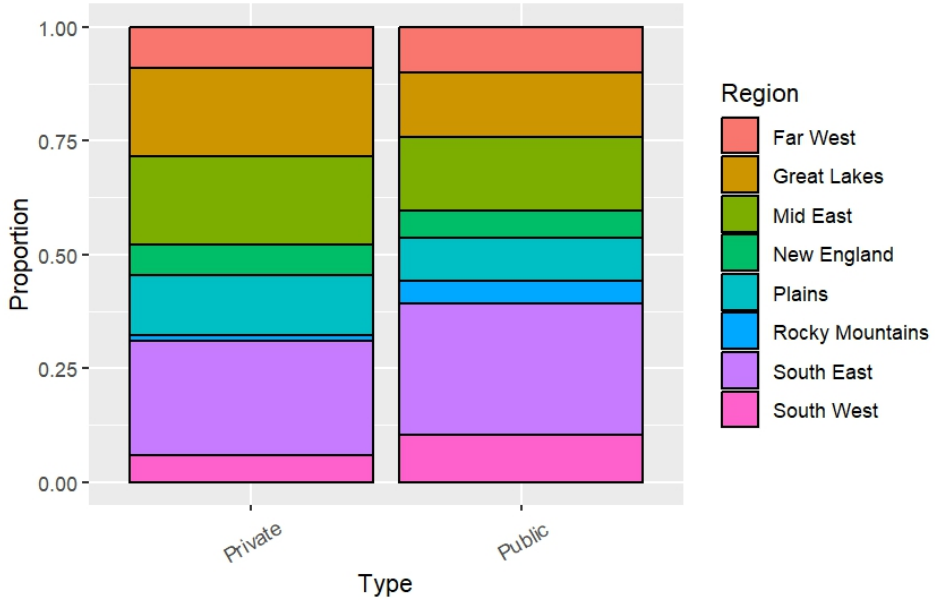




Order of Variables

When we are looking at 2+ variables in a chart, switching the order we display variables can help us think about relationships differently





- of course we need to take care it doesn't become cluttered!